

Линейный регрессионный анализ

Проверка предположений и анализ
результатов

Лекция 7

Пример: вес новорожденного

Файл `birthweight_reduced.csv` содержит данные о 42 новорожденных и их родителях. Нас интересует как зависит вес новорожденного (`Birthweight`, указан в фунтах, 1 брит. фунт = 453.59 грамма) от продолжительности беременности (`Gestation`, в неделях), роста и веса матери до беременности (`mheight` и `mrrwt` в дюймах и фунтах соответственно, 1 дюйм = 2.54 см) и от того, курит ли мать (`smoker`: 0 — не курит, 1 — курит)

Источник: <http://www.statstutor.ac.uk/students/topics/r/statistical-analyses-using-r/>

Чтение и очистка от пропусков

```
babyData = read.csv('birthweight_reduced.csv')  
head(babyData, n=3)
```

```
##      id headcircumference length Birthweight Gestation smoker motherage  
## 1 1313                12      17          5.8         33        0         24  
## 2  431                12      19          4.2         33        1         20  
## 3  808                13      19          6.4         34        0         26  
##  mnocig mheight mppwt fage fedyrs fnocig fheight lowbwt mage35  
## 1      0      58    99  26    16      0      66      1      0  
## 2      7      63   109  20    10     35     71      1      0  
## 3      0      65   140  25    12     25     69      0      0  
##  LowBirthWeight  
## 1              Low  
## 2              Low  
## 3             Normal
```

```
dim(babyData)
```

```
## [1] 42 17
```

```
babyData = na.omit(babyData); dim(babyData)
```

```
## [1] 42 17
```

smoker — это фактор

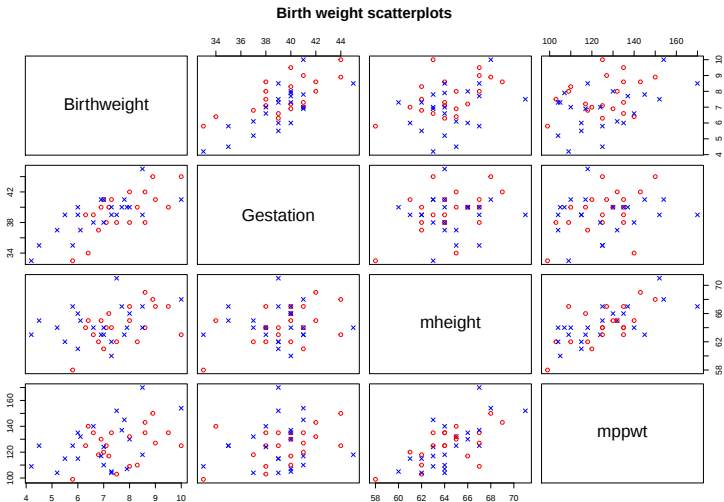
```
class(babyData$smoker)
```

```
## [1] "integer"
```

```
babyData$smoker<-factor(babyData$smoker,  
                        labels=c('Non-smoker', 'Smoker'))
```

Проверка на мультиколлинеарность 1

```
attach(babyData)
pairs(~Birthweight+Gestation+mheight+mppwt,
      main='Birth weight scatterplots',
      col=c('red','blue')[smoker], pch=c(1,4)[smoker])
```



Проверка на мультиколлинеарность 2

Матрица корреляции

```
round(cor(cbind(Birthweight,Gestation,mppwt,mheight)),2)
```

##	Birthweight	Gestation	mppwt	mheight
## Birthweight	1.00	0.71	0.39	0.37
## Gestation	0.71	1.00	0.25	0.23
## mppwt	0.39	0.25	1.00	0.67
## mheight	0.37	0.23	0.67	1.00

Проверка на мультиколлинеарность 3: tolerance

R_j^2 — коэффициент детерминации для регрессии j -го предиктора на остальные предикторы.

$$tolerance = 1 - R_j^2$$

Пороговые значения:

- ▶ 0.2 — допустимо
- ▶ 0.1 — опасно!

Проверка на мультиколлинеарность 3: VIF

VIF (Variance Inflation Factors) = $1/\text{tolerance}$

- ▶ от 1 до 5 — хорошо
- ▶ от 5 до 10 — внимание, опасно!
- ▶ 10+ — данный предиктор существенно зависит от других предикторов

Вычисление VIF в R

```
library(car)  
vif(lm(Birthweight~Gestation+smoker+mheight+mppwt))
```

```
## Gestation      smoker      mheight      mppwt  
##  1.087246  1.013175  1.838976  1.852606
```

```
vif(lm(Birthweight~Gestation+smoker+mppwt))
```

```
## Gestation      smoker      mppwt  
##  1.077986  1.010494  1.068478
```

```
vif(itog1) # Для недвижимости в Альбукерке
```

```
      SQFT      AGE      FEATS      NE      CUST      COR      TAX  
6.197432 1.692336 1.459148 1.374965 1.385906 1.108327 6.476598
```

Считаем VIF'ы в пакете usdm

```
library(usdm)
independents<-data.frame(cbind(Gestation,smoker,mppwt))
vif(independents)
```

	Variables	VIF
1	Gestation	1.077986
2	smoker	1.010494
3	mppwt	1.068478

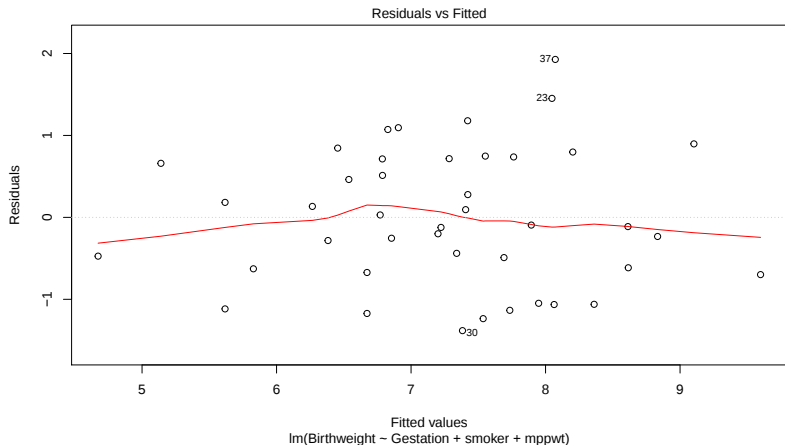
Строим регрессионную модель на основе независимых предикторов

```
reg1<-lm(Birthweight~Gestation+smoker+mppwt)
summary(reg1)
```

```
##
## Call:
## lm(formula = Birthweight ~ Gestation + smoker + mppwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.3824 -0.6246 -0.1033  0.7158  1.9276
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.164740   2.107344  -3.400  0.0016 **
## Gestation      0.313421   0.052887   5.926 7.19e-07 ***
## smokerSmoker  -0.665279   0.267762  -2.485  0.0175 *
## mppwt          0.019819   0.008764   2.261  0.0295 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8622 on 38 degrees of freedom
## Multiple R-squared:  0.6104, Adjusted R-squared:  0.5796
## F-statistic: 19.84 on 3 and 38 DF,  p-value: 6.635e-08
```

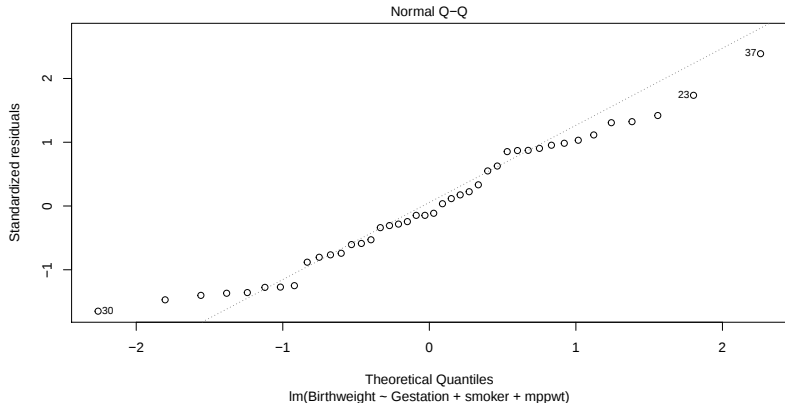
Проверка предположений: нормальность, гомоскедастичность и несмещенность оценок

```
plot(reg1, which = 1)
```



Проверка предположений: нормальность

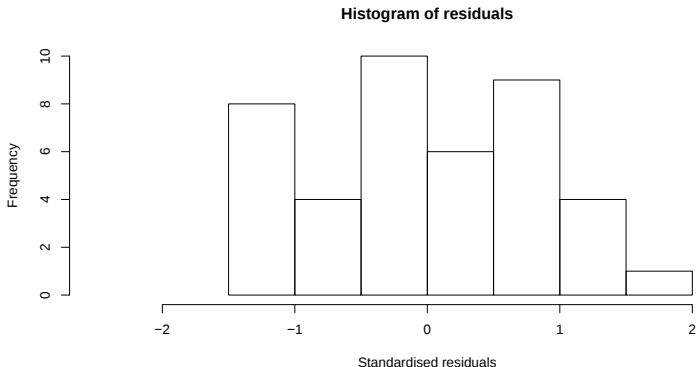
```
plot(reg1, which = 2)
```



О нормальных графиках КК (QQ plot): <https://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot>

Гистограмма для проверки нормальности

```
hist(resid(reg1), xlim = range(c(-2.5,2.5)),  
     main='Histogram of residuals',  
     xlab='Standardised residuals',ylab='Frequency')
```



Мы ожидаем увидеть 95% стандартизованных остатков в диапазоне ± 1.96

Экстремальные значения

Если более 5% остатков выходит за пределы диапазона ± 1.96 или присутствуют экстремальные значения, выходящие из ± 3 , то выполните регрессию с экстремальными значениями и без них, чтобы увидеть насколько сильно присутствие таких наблюдений изменяет коэффициенты модели.

Как выглядят проблемы

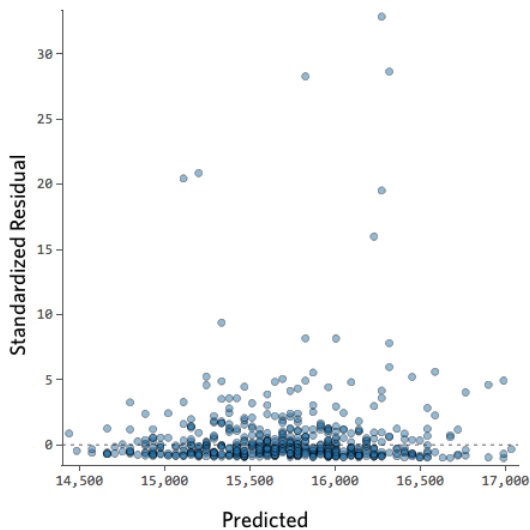


Figure 1:

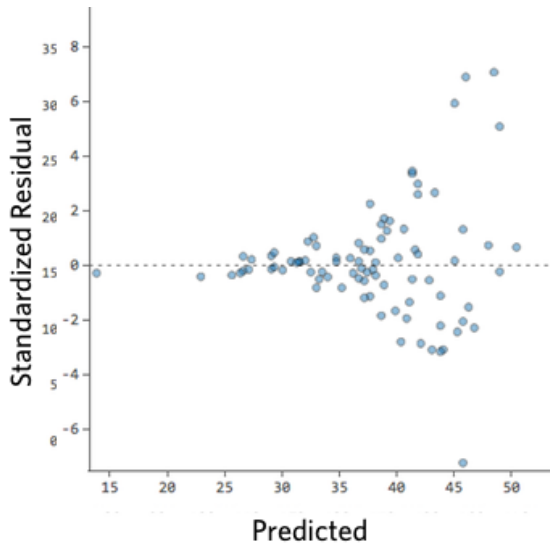


Figure 2:

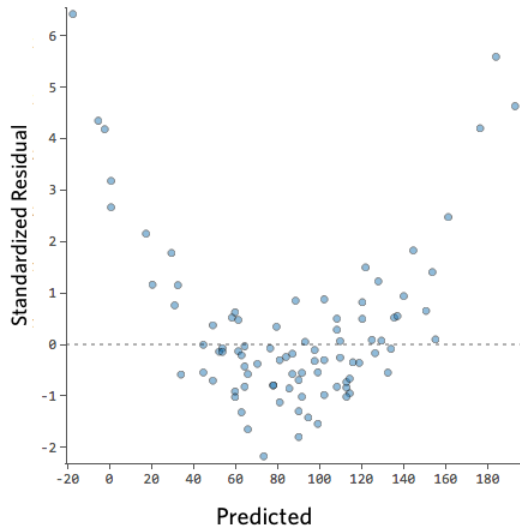


Figure 3:

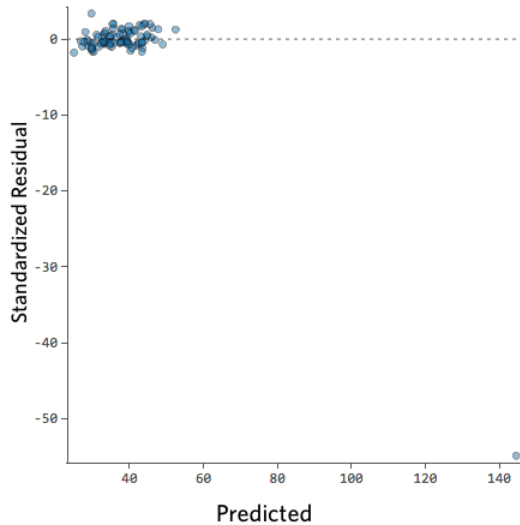


Figure 4:

Проверка на выбросы: расстояние Кука

Расстояние Кука (Cook's distance) для наблюдения показывает, насколько сильно изменится прогнозируемое значение $\hat{Y}$, если удалить это наблюдение из выборки. Иными словами, расстояние Кука характеризует влияние наблюдения.

Какие наблюдения считать влиятельными?

Те, для которых расстояние Кука D_i :

- ▶ > 1
- ▶ $> 4/N$, где N — число наблюдений (для параноиков).

Проверка на выбросы: рычаг

Рычаг (leverage) i -го наблюдения равен $h_{ii} = [H]_{ii}$.

Как понять, что такое рычаг в статистике:

<https://stats.stackexchange.com/questions/58141/interpreting-plot-lm>

Рассмотрим графики остатков для 4-х наборов данных:

1. все хорошо
2. есть точка с длинным рычагом, но малым остатком
3. есть точка с коротким рычагом и большим остатком
4. есть точка с длинным рычагом и большим остатком

```
set.seed(20)
```

```
x1 = rnorm(20, mean=20, sd=3)
```

```
y1 = 5 + .5*x1 + rnorm(20)
```

```
x2 = c(x1, 30);      y2 = c(y1, 20.8)
```

```
x3 = c(x1, 19.44);   y3 = c(y1, 20.8)
```

```
x4 = c(x1, 30);      y4 = c(y1, 10)
```

Все хорошо

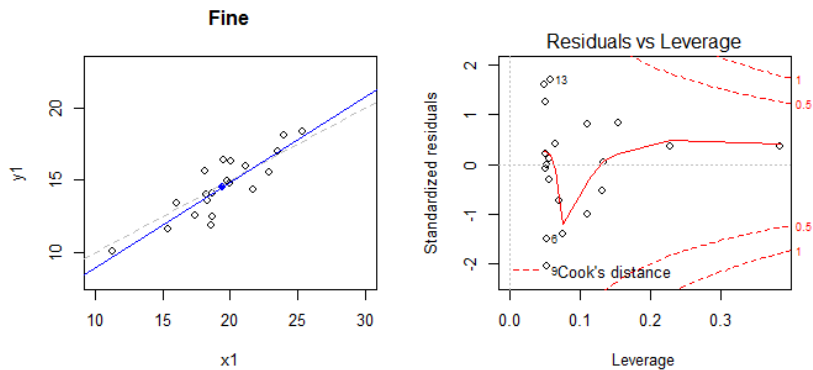


Figure 5:

Точка с длинным рычагом и малым остатком

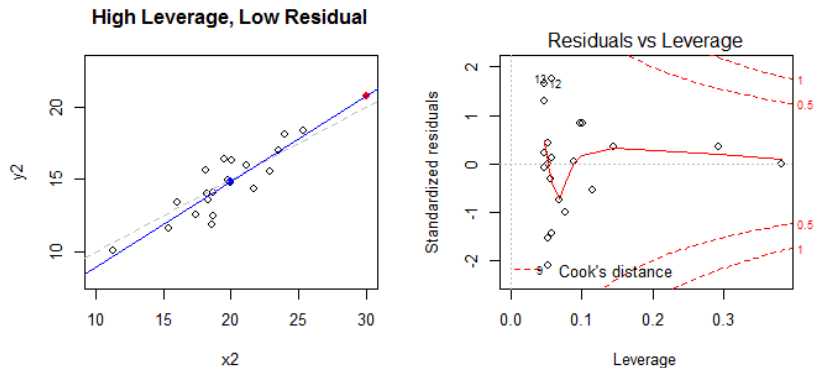


Figure 6:

Точка с коротким рычагом и большим остатком

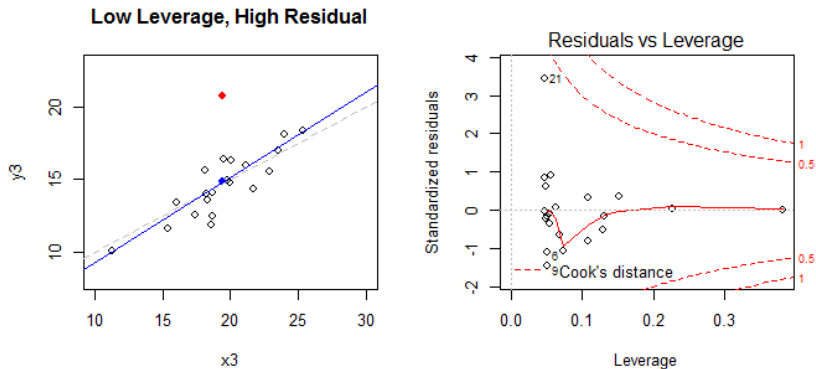


Figure 7:

Точка с длинным рычагом и большим остатком

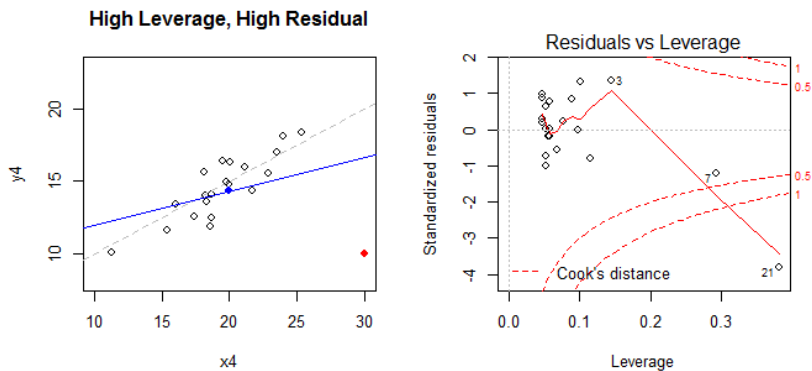
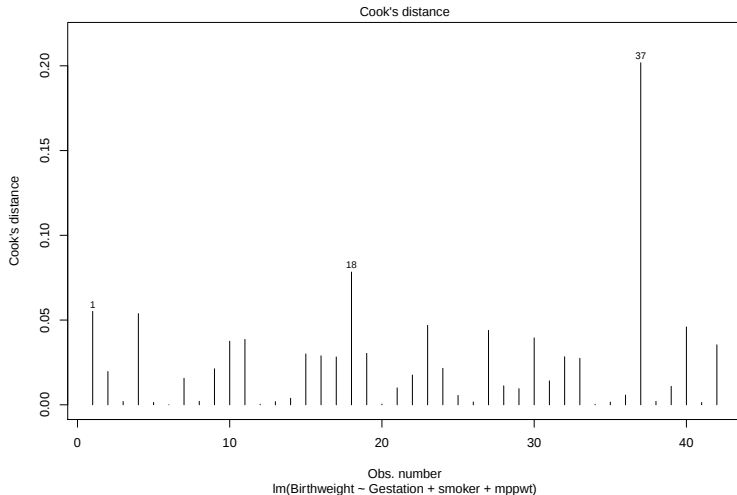


Figure 8:

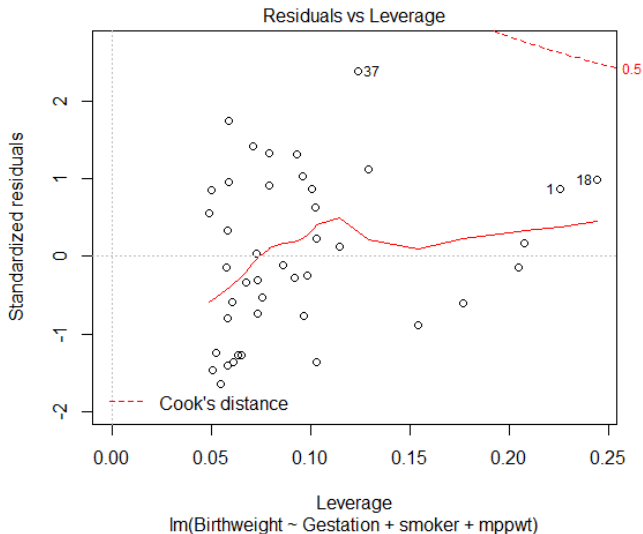
Возвращаемся к исследованию веса новорожденных...

```
plot(reg1, which = 4)
```



4-й график из серии plot()

```
plot(reg1, which = 5)
```



Автокорреляция остатков

Одним из предположений относительно регрессии является независимость наблюдений, которая выражается в отсутствии корреляции между остатками. Однако, если наблюдается процесс развивающийся во времени, то вполне возможно, что последовательные наблюдения будут связаны друг с другом.

Тест Дурбина-Ватсона (Durbin-Watson test) проверяет отсутствие автокорреляции. Если автокорреляции нет (т.е. последовательные наблюдения не взаимосвязаны), то статистика Дурбина-Ватсона будет находится в пределах от 1.5 до 2.5. p -значение обычно принимается равным 0.05.

Тест Дурбина-Ватсона

```
library(car)  
dwt(reg1)
```

```
lag Autocorrelation D-W Statistic p-value  
1      -0.0774652      2.122239  0.782  
Alternative hypothesis: rho != 0
```

Предположение об отсутствии автокорреляции обычно нарушается для временных рядов.

Выводы на основе регрессионной модели

Множественная линейная регрессия проводилась в целях изучения взаимосвязи между весом новорожденного и 1) продолжительностью беременности (в неделях), 2) весом матери до беременности и 3) курением матери в течение беременности.

Показано наличие значимой взаимосвязи между продолжительностью беременности (гестационным возрастом) и весом новорожденного ($p < 0.001$), курением матери и весом новорожденного ($p = 0.017$), а также весом матери до беременности и весом новорожденного ($p = 0.03$).

Отмечено увеличение веса новорожденного на 0.313 фунта за каждую дополнительную неделю гестационного возраста.

На каждый дополнительный фунт веса матери, вес новорожденного увеличивается на 0.02 фунта.

Новорожденные у курящих матерей весят на 0.665 фунта меньше, чем у некурящих.

Скорректированный коэффициент детерминации (adjusted R-squared) равен 0.58, следовательно 58% изменений веса новорожденных может быть объяснено регрессионной моделью, включающей гестационный возраст, вес и курение матери.

Теорема Гаусса-Маркова.

Схема регрессионного анализа

Для того чтобы полученные по МНК оценки обладали некоторым полезными статистическими обладали некоторыми полезными статистическими свойствами, необходимо выполнение ряда условий относительно оцениваемой модели, называемых условиями Гаусса-Маркова.

Теорема Гаусса-Маркова

Рассматривается модель парной регрессии, в которой наблюдения Y связаны с X следующей зависимостью:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i.$$

На основе n наблюдений оценивается уравнение регрессии

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i.$$

Тогда, если данные обладают следующими свойствами:

1. Модель данных правильно специфицирована;
2. Все X_i детерминированы и не все равны между собой;
3. Ошибки не носят систематического характера, то есть $\mathbb{E}(\varepsilon_i) = 0 \ \forall i$;
4. Дисперсия ошибок одинакова и равна некоторой σ^2 ;
5. Ошибки некоррелированы, то есть $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \ \forall i, j$;

— то оценки полученные методом наименьших квадратов оптимальны в классе линейных несмещённых оценок.

Пояснения

- ▶ Эффективность оценки означает, что она обладает наименьшей дисперсией.
- ▶ Оценка линейна по наблюдениям Y .
- ▶ Несмещённость оценки означает, что её математическое ожидание равно истинному значению.

Условия Гаусса-Маркова 1

Спецификация модели

$$y_i = b_0 \cdot x_{i0} + b_1 \cdot x_{i1} + b_2 \cdot x_{i2} + \dots + b_k \cdot x_{ik} + \varepsilon_i, \quad x_{i0} = 1, \quad i = 1, 2, \dots, n.$$

1. Если модель линейна по параметрам b_i , то спецификация корректна.
2. Отсутствие коллинеарности.

Условия Гаусса-Маркова 2

Все X_i детерминированы и не все равны между собой.

Если все X_i равны между собой, то $X_i = \bar{X}$, и в уравнении оценки коэффициента наклона прямой в линейной модели в знаменателе будет ноль, из-за чего будет невозможно оценить коэффициенты b_i .

Условия Гаусса-Маркова 3

Ошибки не носят систематического характера:

$$\mathbb{E}(\varepsilon_i) = 0 \quad \forall i$$

Случайный член может быть иногда положительным, иногда отрицательным, но он не должен иметь систематического смещения ни в каком из двух возможных направлений.

Если уравнение регрессии включает постоянный член b_0 , то это условие выполняется автоматически, так как постоянный член отражает любую систематическую постоянную составляющую в Y , которую не учитывают объясняющие переменные, входящие в уравнение регрессии.

Условия Гаусса-Маркова 4

Дисперсия ошибок одинакова.

Одинаковость дисперсии ошибок принято называть **гомоскедастичностью**.

Если это условие не выполняется, то коэффициенты регрессии, найденные по методу наименьших квадратов, будут неэффективны

Более эффективные результаты будут получаться путём применения модифицированных методов оценивания (взвешенный МНК или оценка ковариационной матрицы по формуле Уайта или Дэвидсона—Маккинона).

Условия Гаусса-Маркова 5

Ошибки ε_i распределены независимо от ε_j при $i \neq j$.

Это условие предполагает отсутствие систематической связи между значениями случайного члена в любых двух наблюдениях. Если один случайный член велик и положителен в одном направлении, не должно быть систематической тенденции к тому, что он будет таким же великим и положительным (то же можно сказать и о малых, и об отрицательных остатках).

Обычно не выполняется в случае временных рядов!

При невыполнении этого условия оценки, полученные по методу наименьших квадратов, будут неэффективны.

Простая линейная регрессия

```
x <- c(10, 20, 30, 40, 50)
y <- c(4, 22, 44, 60, 82)
```

На основе $n = 5$ наблюдений строим уравнение регрессии

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X.$$

которое является оценкой для

$$Y = b_0 + b_1 X + \varepsilon.$$

- ▶ X — предиктор
- ▶ Y — отклик
- ▶ b_0, b_1 — коэффициенты регрессии

$\varepsilon = Y - \hat{Y}$ — ошибка (остаток, погрешность, невязка) равна разности между наблюдаемым и прогнозируемым значениями отклика.

Residual = Observed - Predicted

Этапы линейной регрессии в R

1. Собрать (получить) выборку числовых данных о связи предикторов с откликом.
2. Записать формулу, предположительно связывающую предикторы и отклик. Поместить эту формулу в функцию `lm()`.
3. Проанализировать качество регрессионной модели при помощи `summary()`.
4. По результатам выполнения `lm()` найти коэффициенты регрессии. Записать с их помощью уравнение регрессии $Y = b_0 + b_1 X$.
5. Для прогнозирования новых значений отклика использовать функцию `predict()`. Определить доверительный интервал прогноза.

`lm()` — функция для подгонки линейных моделей

```
lm(Y ~ model, data)
```

Формула:

```
Y ~ op1 X1 op2 X2 op3 X3 ...
```

представляет собой объект класса `formula`.

`data` — таблица данных

Функция `lm` возвращает линейную регрессионную модель.

Формулы

“Линейная модель” означает линейность относительно коэффициентов регрессии b_i

- ▶ $y = b_0 + b_1x + b_2x^2$ — линейная модель.
- ▶ $y = b_0x^{b_1}$ — нелинейная модель.

Примеры

- ▶ $Y \sim A$ — прямая со свободным членом b_0 , заданным неявно

$$Y = b_0 + b_1A$$

- ▶ $Y \sim -1 + A$ — прямая без свободного члена; будет проходить через (0,0)

$$Y = b_1A$$

- ▶ $Y \sim A + I(A^2)$ — полином; внутри функции $I()$ можно задавать операции, которые трактуются в обычном для математики смысле

$$Y = b_0 + b_1A + b_2A^2$$

Примеры формул

- ▶ $Y \sim A + B$ — модель 1-го порядка, в которой A и B не взаимодействуют

$$Y = b_0 + b_1A + b_2B$$

- ▶ $Y \sim A:B$ — модель, содержащая только взаимодействие между A и B (1-го порядка)

$$Y = b_0 + b_1AB$$

- ▶ $Y \sim A*B$ — полная модель 1-го порядка, учитывающая как A и B , так и взаимодействие между ними (эквивалент: $Y \sim A + B + A:B$)

$$Y = b_0 + b_1A + b_2B + b_3AB$$

- ▶ $Y \sim (A + B + C)^2$ — модель включает все эффекты 1-го порядка и все взаимодействия вплоть до n -го порядка, где n задается показателем в $(\dots)^n$ (эквивалент: $Y \sim A*B*C + A:B:C$)

$$Y = b_0 + b_1A + b_2B + b_3C + b_4AB + b_5AC + b_6BC$$

```
lm.y <- lm(y ~ x)
summary(lm.y)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      1      2      3      4      5
##  0.4 -1.0  1.6 -1.8  0.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.80000    1.66933   -9.465  0.0025 **
## x             1.94000    0.05033  38.544 3.84e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.592 on 3 degrees of freedom
## Multiple R-squared:  0.998, Adjusted R-squared:  0.9973
## F-statistic: 1486 on 1 and 3 DF, p-value: 3.842e-05
```

summary()

- ▶ Residuals — остатки.
- ▶ Оценки коэффициентов модели (Estimate), их стандартные отклонения (Std Error), t-значения и вероятности нулевой гипотезы, что коэффициент равен нулю.
- ▶ стандартное отклонение регрессии (Residual standard error).
- ▶ коэффициенты детерминации.
- ▶ результаты F-теста нулевой гипотезы об одновременном равенстве нулю всех коэффициентов регрессионной модели.

Полезные функции

```
coef(lm.y)    # коэффициенты модели
```

```
## (Intercept)          x  
##      -15.80         1.94
```

```
resid(lm.y)   # остатки
```

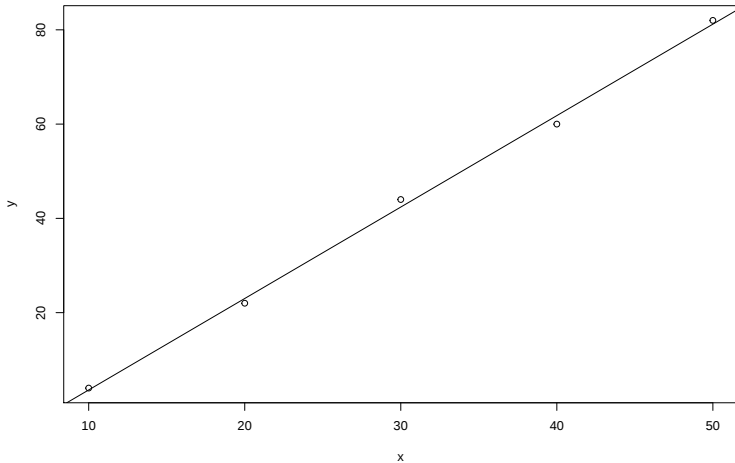
```
##      1      2      3      4      5  
##  0.4 -1.0  1.6 -1.8  0.8
```

```
fitted(lm.y)  # подогнанные значения Y
```

```
##      1      2      3      4      5  
##  3.6 23.0 42.4 61.8 81.2
```

$$y_h = -15.8 + 1.94 \cdot x_h$$

```
plot(x, y)  
abline(lm.y)
```



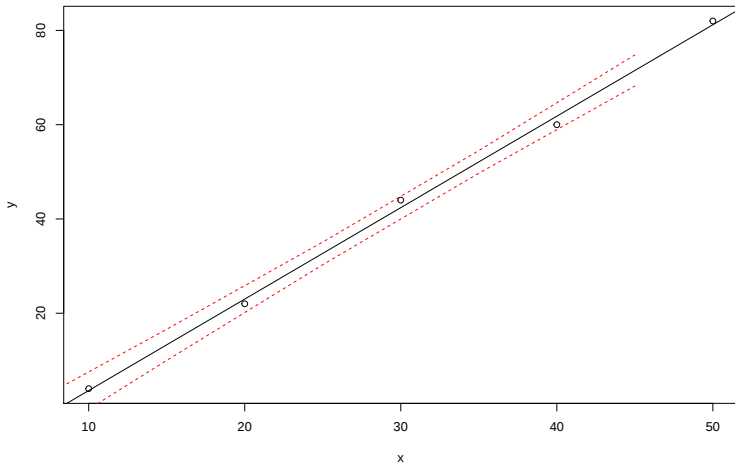
Прогнозирование и доверительный интервал

```
predict(model, data.frame(новые данные),  
        level = 0.95, interval = "confidence")
```

```
x.new = c(5,15,25,35,45)  
preds <- predict(lm.y, data.frame(x = x.new),  
                 level = 0.95, interval = "confidence")
```

```
##      fit      lwr      upr  
## 1 -6.1 -10.700809 -1.499191  
## 2 13.3   9.997813 16.602187  
## 3 32.7  30.297306 35.102694  
## 4 52.1  49.697306 54.502694  
## 5 71.5  68.197813 74.802187
```

```
plot(x, y)
abline(lm.y)
lines(x.new, preds[,3], lty = 'dashed', col = 'red')
lines(x.new, preds[,2], lty = 'dashed', col = 'red')
```



Квадратичная аппроксимация

```
# Данные Галилея
height = c(100, 200, 300, 450, 600, 800, 1000)
distance = c(253, 337, 395, 451, 495, 534, 574)
# Модель в форме квадратичного полинома
lm.r = lm(distance ~ height + I(height^2))
# Создадим высоты для прогноза
newh = seq(100, 1000, 10)
# Вычислим расстояния для каждой из новых высот
fit = 200.211950 + 0.706182*newh - 0.000341*newh^2
```

Квадратичная аппроксимация: график

```
plot(height, distance) # исходные данные  
lines(newh, fit, lty=1) # показать результаты подгонки
```

